OPEN ACCESS

Ethics of the algorithmic prediction of goal of care preferences: from theory to practice

Andrea Ferrario ,[1,2] Sophie Gloeckler ,[3] Nikola Biller-Andorno [3]

[1]ETH Zurich, Zurich, Switzerland
[2]Mobiliar Lab for Analytics at ETH, ETH Zurich, Zurich, Switzerland
[3]Institute of Biomedical Ethics and History of Medicine (IBME), University of Zurich, Zurich, Switzerland

Correspondence to
Dr Andrea Ferrario, ETH Zurich, 8092 Zurich, Switzerland; aferrario@ethz.ch

Received 29 April 2022
Accepted 19 October 2022
Published Online First 8 November 2022

ABSTRACT

Artificial intelligence (AI) systems are quickly gaining ground in healthcare and clinical decision-making. However, it is still unclear in what way AI can or should support decision-making that is based on incapacitated patients' values and goals of care, which often requires input from clinicians and loved ones. Although the use of algorithms to predict patients' most likely preferred treatment has been discussed in the medical ethics literature, no example has been realised in clinical practice. This is due, arguably, to the lack of a structured approach to the epistemological, ethical and pragmatic challenges arising from the design and use of such algorithms. The present paper offers a new perspective on the problem by suggesting that preference predicting AIs be viewed as sociotechnical systems with distinctive life-cycles. We explore how both known and novel challenges map onto the different stages of development, highlighting interdisciplinary strategies for their resolution.

INTRODUCTION

In clinical settings, preference-sensitive decisions regarding diagnostic and treatment options often need to be determined when patients are incapacitated and unable to make such choices for themselves. Advance directives, which allow patients to declare preferences regarding future care while still competent, are often lacking or inconclusive. In such cases, surrogate decision-makers or next-of-kin help establish what the patient would have presumably wanted. Ideally, their insight into the patient's preferences along with clinicians' expertise come together in a process of shared decision-making that aims to define the best possible treatment for the patient. However, effectively determining what is in the patient's best interest in accordance with their values and goals on their behalf can be challenging in practice.

This is especially true in the case of severe and unexpected events that necessitate timely intensive care interventions. In these situations, providing goal concordant care, that is, care that corresponds to patient preferences, requires clarity around whether patients would prefer palliative approaches or lifesaving interventions depending on the likelihood and extent of cognitive or motor deficits, survival rates and the burden of the treatment itself. Clinicians must balance a range of ethically challenging and complex demands, including providing treatment consistent with the patient's presumed preferences and values; supporting the patient's family and loved ones; ensuring timely decision-making and overseeing clinical care.[1] However, as the literature has shown, there is often a considerable gap between the care that is received and the care patients would have wanted.[2 3] Moreover, available instruments to record goal of care preferences, such as advance directives, still do not provide the guidance clinicians often need to identify patient preferences.[4 5] Moreover, those close to the patient tend to be part of the decision-making process, but research has shown that surrogates and next-of-kin are often limited in their ability to accurately predict their loved ones' preferred care[6 7] and often suffer emotional distress in the face of such decision-making.[1]

Researchers have proposed to address the aforementioned challenges by designing and implementing algorithms that would compute the most likely preferred treatment of the incapacitated patient.[8–10] To the best of our knowledge, the first of these proposals dates back to 2010, when Rid and Wendler introduced the idea of using patients' sociodemographic data to predict preferred treatment options.[1] Since 2010, other algorithmic proposals have been made: they all share the idea that algorithms can learn patterns in data that correlate individual-level information to preferred treatments for a variety of clinical scenarios. Despite preliminary evidence suggesting that patients, surrogates and clinicians may respond positively to the use of these algorithms in clinical practice,[9 11 12] and the successful use of artificial intelligence methods in other medical applications,[13] a decade after Rid and Wendler's proposal, no working example of systems to predict patient's preferences in clinical practice has yet emerged.

The reasons, we argue, are manifold. First, the idea of using algorithms to predict patient's preferred treatments is fraught with theoretical challenges that pertain to epistemology and ethics. Some of these challenges relate to patient autonomy,[14–16] the difficulty of avoiding bias and the importance of addressing explainability given the 'black box' nature of many artificial intelligence (AI) algorithms.[8] Second, the design, development and use of these algorithms need to address pragmatic and human–computer interaction (HCI) driven challenges, such as safety, reliability and adequate testing to foster usability and trust among clinicians.[17 18] Lastly, the literature mostly addresses the aforementioned theoretical and pragmatic challenges separately, missing the opportunity to promote a more comprehensive discussion that considers the relationship between the two and would be better positioned to address such interdisciplinary concerns.

The goal of the present paper is to support both the design of algorithms to predict patient preferences and their implementation in clinical practice

Linked

► http://dx.doi.org/10.1136/jme-2023-108945

Check for updates

To cite: Ferrario A, Gloeckler S, Biller-Andorno N. *J Med Ethics* 2023;**49**:165–174.

by centring questions of ethics and considering a life-cycle perspective of the clinical AI systems in which these algorithms ought to be embedded.[i] This view moves away from the existing algorithm-centred perspective, instead recognising that an AI in clinical applications is more than an algorithm and a data flow: it is a system interacting with stakeholders according to the different stages of its life-cycle embedded in a particular socio-technical context.[19] Focusing on the particularities of all stages of the AI system life-cycle, the aim is to capture the relevant interactions with stakeholders without limiting the discussion exclusively to the challenges stemming either from conceptualisation (eg, data collection) or from the integration in clinical decision-making (eg, during the possible interactions with surrogates). In fact, we consider the before, the during and the after of the AI system use in clinical practice and the relation between these phases. To do so, we break down the life-cycle of the clinical AI system in five sequential steps, highlighting theoretical and pragmatic challenges sequentially. To guide our exposition, we focus on a use case of relevance for goal concordant care, namely decision-making in the intensive care unit (ICU). However, the proposed approach can be generalised to other use cases of clinical relevance, such as incapacitation in cases of severe dementia.

The main benefits of our approach are the following. First, it allows us to map systematically existing epistemological, ethical and pragmatic challenges, such as the necessity to respect patient autonomy or to promote appropriate explanations of the predicted treatments, onto the different steps of the AI system life-cycle. Second, it allows us to identify and discuss novel challenges, such as those tied to the methodology of studies retrospectively evaluating critical care. Finally, it provides actionable recommendations to support the design and use of these systems in concordance with normative design requirements. As a result, we argue, our approach could promote an interdisciplinary dialogue between ethicists, clinicians and computer scientists focusing on the still-standing challenges that arise from the concept of AI-assisted prediction of patient preferences and finally supporting the integration of such systems into clinical practice.

The plan of the paper is as follows. First, we introduce a clinical use case that will guide the discussions throughout the paper. Then, we discuss the current state of debate on algorithmic prediction of patient preferences. Finally, we introduce our approach. We then present our conclusions.

A USE CASE FOR PREDICTING PATIENT PREFERENCES: DECISION-MAKING IN THE ICU

Decisions in the ICU must be made quickly due to the severity of the fast-evolving clinical presentation and the time-sensitive nature of many of the interventions. In addition, in the ICU, patients are often incapacitated, such as in the case of haemorrhagic stroke, and cannot directly consent to or refuse interventions. While clinicians primarily seek to follow the will of the patient, including in the form of respect for any existing advance directive, input may be needed from loved ones and healthcare professionals if the patient is incapacitated and an advance directive is either not available or not readily applicable to the presenting scenario.

De facto, ethical, institutional, social and pragmatic constraints, such as time pressure and whether there is any indication of relevant care preferences, may influence which treatments are administered and hamper clinicians' efforts to provide patient-centred, goal concordant care. The above specificities make the ICU a paradigmatic case where methods of evidence-based technological support to promote shared decision-making can be implemented and where well-considered innovation can have significant impact on the quality of care.[20]

Advance directives, a form through which users leave written instructions outlining their future healthcare preferences, play an important role in determining care for incapacitated patients. Although research has shown that patients desire advance directives and that surrogate decision-makers find them helpful,[21 22] completion rates remain generally low.[4] Moreover, advance directives have significant shortcomings, such as content that can be difficult to understand or have limited relevance to the concerns of certain ethnic and social groups.[5] Finally, advance directives may be unavailable when needed, even if completed, or may fail to provide relevant guidance for the scenario at hand.[23]

If an incapacitated patient does not have an advance directive, then surrogate decision-makers or next-of-kin help to identify the presumed will of the patient, drawing on past conversations or comparable life choices. However, literature shows that these stand-in decision-makers have a limited ability to accurately predict the goal of care preferences of their loved ones.[6 24] Moreover, making such life and death decisions for another often generates significant emotional distress[1] and needed guidance from ICU clinicians can be lacking.[25] Finally, surrogate decision-makers may not be available for consultation when needed, leaving the burden of deciding whether or not to pursue certain treatments to clinicians.

In the remainder of this paper, we will use decision-making in the ICU as a use case to clarify our discussions on the prediction of patients' preferred treatments with the use of technology. However, our considerations apply to other clinical settings, such as the emergency and palliative care unit, and can be generalised to other scenarios of incapacitation, such as cases of severe dementia.

ALGORITHM-AIDED PREDICTION OF PATIENT PREFERENCES

In light of the difficulties arising from the shortcomings of advance directives and the limitations of stand-in decision makers, researchers have suggested the potential value of designing algorithms that predict patients' most likely preferences to support clinical decision-making for those unable to make or express decisions on their own behalf.[1 8–10 26 ii] Rid and Wendler's patient preference predictor (PPP) is an early example of such an algorithm.[1] Once fed on appropriate data, the algorithms would compute a preferred treatment for a pre-selected variety of clinical interventions. More recent contributions suggest the use of AI rather than algorithms that use a pre-defined set of criteria,[iii] projecting use cases such as a Do-Not-Attempt-to-Resuscitate Predictor.[9] As input data, demographic information that influences care preferences,

[i]The life-cycle is a representation of the process of design, development, deployment and maintenance of a system, such as an AI. In this work, the life-cycle of the AI predicting goal of care preferences comprises five steps: (1) 'Data Collection', (2) 'Mmodelling', (3) 'System Design', (4) 'Deployment' and (5) 'Evaluation'.

[ii]In this context, an algorithm is a set of deterministic of probabilistic rules run by computer systems to predict an output of interest, given input data.

[iii]An example is the population-based rule stating that *any* incapacitated patient will prefer life-saving treatment 'when there is at least a 1% chance, following the intervention, that the patient will reach a health state which includes the ability to reason, remember and communicate'.[24]

such as age, gender, marital status, health condition and previous healthcare experiences, could be gathered alongside more general aspects of the person's known goals and values to produce either a population-based or individual-based prediction.

The working hypothesis states that an accurate algorithm increases the chance that patients receive treatment consistent with their preferences.[14] The hypothesis is supported by empirical evidence showing that an average predictor is epistemically comparable to surrogates and next-of-kin in predicting preferences.[24 27 28]

The use of AI methods, such as machine learning (ML), that have proven to deliver high performance in multiple clinical use cases,[13] would allow for the generation of personalised predictions of goal of care preferences, instead. It is likely that these personalised predictions will be more accurate (on average) than those of surrogates or next-of-kin.[1]

Moreover, high-performing AI-based algorithms have the potential to reduce the distress of those asked to make treatment decisions for an incapacitated loved one[29] by aiding them in incorporating patients' values and preferences into considerations in an evidence-based way.[20]

However, although preliminary studies show that clinicians, patients and surrogate decision-makers are in favour of the introduction of algorithms to support decision-making,[9 11 12 iv] and despite the increasing use of AI in healthcare,[13] no such algorithm has been integrated into clinical practice.

Actually, the proposal of using algorithms to predict patient preferences is not free from criticism pointing to unresolved theoretical and practical challenges. However, the former, which pertain to the epistemology and ethics of algorithms, rarely inform the latter, which discuss limitations in the design and possible use of algorithms in clinical practice. For example, some authors highlight the necessity of tackling bias in data collection; promoting the use of methods to increase explainability of opaque AI algorithms; and providing transparent, secure and reliable infrastructure for their use.[8] Others focus on ethical challenges, highlighting how such algorithms could potentially endanger patient autonomy and the importance of addressing the perceived acceptability of such tools on the part of potential users.[16 26 30] These criticisms often emphasise concern over how algorithms might diminish or confound the role of patients and their families as decision-makers. The problem of limiting freedom of choice by using demographic data as predictors of preferred treatments is also a common concern.[16 30] Despite recent efforts,[14] these challenges remain largely unaddressed.

In summary, these discussions capture some of the limitations stemming from algorithmic predictions of patient preferences, with and without AI. Attention typically centres on the algorithm and its input-output data flows at specific stages. This perspective, however, falls short by neglecting exploration of the sociotechnical context of such a decision support system at all stages of its use. An alternative approach would support researchers in responsibly closing the gap between conceptualisation and actual use in clinical practice, providing a structure for discourse around epistemological, ethical and pragmatic challenges and subsequent strategies of resolutions. We defend this proposal in the forthcoming sections.

ALGORITHMS TO PREDICT PATIENT PREFERENCES: A SOCIOTECHNICAL SYSTEM APPROACH

A sociotechnical system is the result of a technical artefact, for example a computer system, in a dialogue with human agents set within the specification of social norms or rules that regulate the interactions.[19] A clinical AI that predicts patient preferences is an example of a sociotechnical system: the technical artefact, that is, a computer system, generates suggested treatments and interacts with a set of human agents (eg, patients, clinicians, nurses and loved ones), from within the set of social norms specified by practices relating to emergency decision-making and the value of respect for patient-centred care. In particular, social norms inform a series of normative requirements deemed necessary for the acceptance and use of AI in clinical practice. Examples include safety, robustness, reliability, privacy, security, transparency, explainability, algorithmic fairness and non-discrimination.[18]

We have previously shown that the algorithm and data flow centred debate on predicting patient preferences has prompted noteworthy research efforts. Still, the articulated theoretical considerations have not yet informed the development of such algorithms or their testing in empirical studies. Vice versa, technical system requirements are not discussed through the lens of epistemology and ethics beyond high-level proposals[8] or discussions limited to specific topics, such as the design of surveys for data collection.[26]

In summary, we argue that a siloed approach falls short of appropriately tackling the complexity of the problem at hand. Therefore, our proposal is to leverage the complexity by promoting a structured discourse that holds a clinical AI preference predictor as a sociotechnical system that evolves over time. By definition, this means considering the technical artefact (ie, the clinical AI of which the evolving algorithm for preference predictions is a part), user interactions, and the social institutions at each relevant point of time within the system life-cycle. To do so, we model the system life-cycle as a sequence of five steps: (1) 'Data Collection', (2) 'Modelling', (3) 'System Design', (4) 'Deployment' and (5) 'Evaluation'. These are presented in figure 1.[v] At each life-cycle step, we highlight theoretical and pragmatic challenges and discuss the relevant interactions with stakeholders, including the use of additional digital tools. The normative requirements for clinical AIs[18] underline all steps of the AI system life-cycle.

Following this approach, we can map the challenges discussed in the literature, such as respecting patient autonomy or fostering the explainability of predictions, into different steps of the AI life-cycle, and therefore, discuss step-specific strategies for resolution. Moreover, the discussion of the steps of the AI life-cycle supports the identification of new challenges that designers, ethicists and clinicians need to address. As a result, our approach provides a theory-driven yet application-oriented roadmap to the implementation of AI systems to predict patient preferences in clinical practice. We collect the steps, their corresponding challenges and our recommendations in figure 2. These are discussed in detail in the forthcoming sections.

[iv]This evidence, together with the possibility of a timely and consistent computation of personalised predictions, the privacy-preserving management of patient's information, and the opportunity of conducting extensive validations of the AI accuracy over time, 'provide compelling reason to pursue future work to evaluate the acceptability and feasibility of using (AI-based) predictions of patients' treatment preferences in practice'.[1]

[v]For the sake of readability, we avoid showing the feedback loops between the different life-cycle phases, such as from 'Modelling' to 'Data Collection' that may occur in practice.

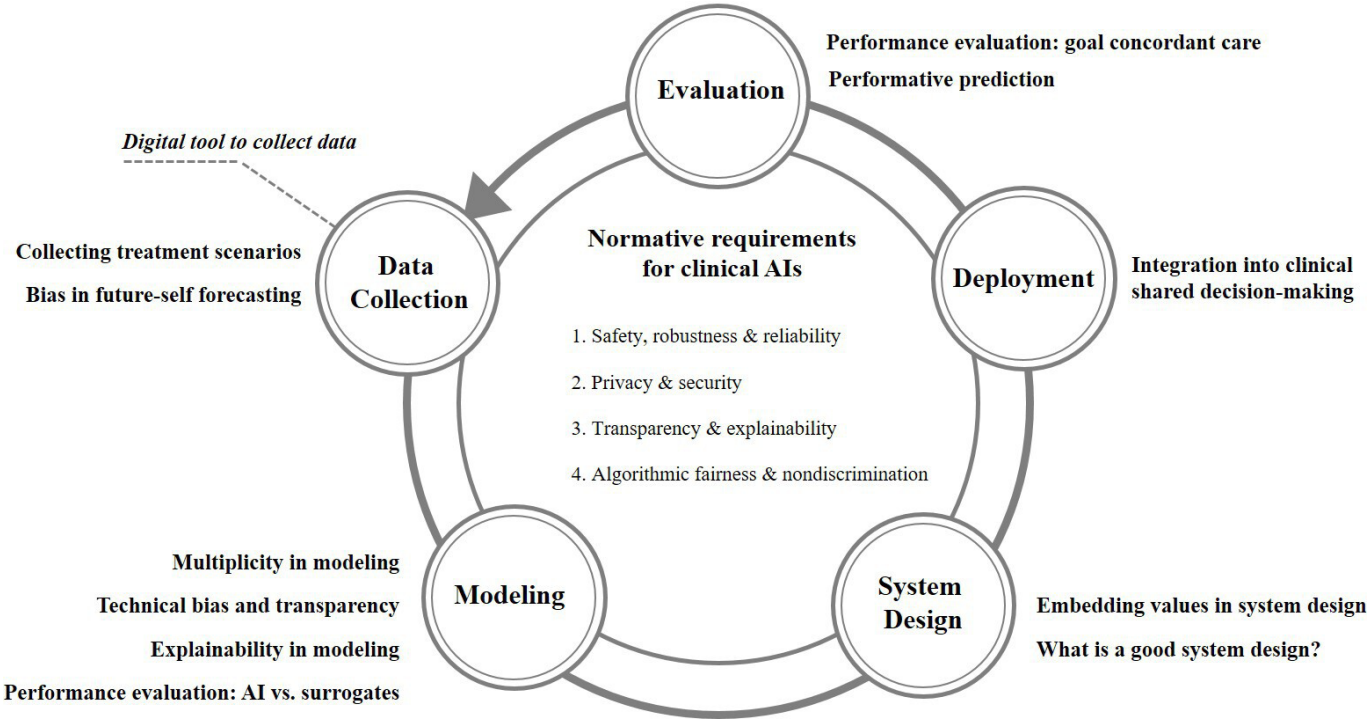

Figure 1 The AI system life-cycle. At each step, we highlight the main theoretical and pragmatic challenges. The normative requirements for clinical AIs underline all steps of the life-cycle. AI, artificial intelligence.

Data collection

Collecting treatment scenarios

When surveying for treatment preferences, the following dimensions must be weighed against one another: the burden of treatment, that is, the hardship endured while being treated; the outcome, that is, the health state and length of life after treatment; and the likelihood of any such outcome.[31] Given this complexity, the number of questions needed to capture someone's preferences represents a challenge for data collection. For example, Rid and Wendler suggest 110–130 questions that may cover 10–20 treatment scenarios,[10] which is, arguably, a rather intensive exercise for any respondent. Few studies have explored the validity of instruments for eliciting input to determine preferred treatments. One example, the WALT instrument, does so in fewer questions, considering multiple aspects of treatment preferences while presenting only six clinical scenarios.[31] Wendler *et al*'s study[11] puts forward only

Life-cycle Steps	Challenges	Recommendations
Data Collection	1. *Collecting treatment scenarios*	1. Collect treatments and outcomes letting responders indicate preferred likelihoods and durations of post-intervention health states
	2. *Bias in future-self forecasting*	2. Development of digital tools to manage bias in future-self forecasting using narratives
Modeling	1. *Multiplicity in modeling*	1. Remove low-performance ML models, aggregate "contiguous" treatment scenarios, possibly refining "Data Collection"
	2. *Technical bias and transparency*	2. Endorse transparency frameworks that include technical bias emerging during "Modeling"
	3. *Explainability in modeling*	3. Use explainability methods to assess feature importance (e.g., Shapley values) and explain treatment predictions (e.g., counterfactuals)
	4. *Performance evaluation: AI vs. surrogates*	4. Use the AI vs. surrogates performance evaluation as prima facie evidence of the AI accuracy over different treatment scenarios, only. Refine "Data Collection" following the outcomes of the evaluation
System Design	1. *Embedding values in system design*	1. Use value-sensitive design ("embedded values") to guide design of the AI system and define its requirements
	2. *What is a good system design?*	2. Use user-centred design to ensure appropriate levels of user experience, usability, the timely use of the AI predictions and its appropriate integration in clinical workflows
Deployment	1. *Integration into clinical shared decision-making*	1.1 Design different integration scenarios for the AI to predict preferences 1.2 Integrate different strategies of resolution (e.g. including clinical ethicist) for disagreements 1.3 Train clinicians in structuring discussions on patient's preferred treatments, providing emotional support and emphasizing shared decision-making
Evaluation	1. *Performance evaluation: goal concordant care*	1. Performance evaluations in different retrospective studies (with clinicians and discharged patients) to measure retrospective agreement
	2. *Performative prediction*	2. Manage the bias induced by performative prediction and affecting "Data Collection" via retrospective studies

Figure 2 The steps of the AI system life-cycle, their challenges and our recommendations. AI, artificial intelligence.

a single-scenario questionnaire in which respondents express their preferences in the case of lost mental capacity due to a car accident. It is still uncertain what is necessary and sufficient for such tools to capture the respondent's position well. In summary, a key step in generating data that preference predicting algorithms could use is to successfully manage the complexity of the variables involved: clinically relevant granularity must be captured, but the number of variables and scenarios should not overwhelm the respondent or interpreter.

One solution we suggest would be to fix an outcome for each treatment of interest and ask respondents to indicate the minimum or maximum likelihood of the outcome at which they would consent to treatment and, where relevant, for what minimum duration following. For example, respondents may be asked to indicate the maximum likelihood of severe cognitive impairment following cardiopulmonary resuscitation they would accept and still desire resuscitation or the minimum number of years of survival following an invasive procedure that would be necessary for them to desire the treatment. Respondents could indicate their choices through a simple sliding scale on a digital interface. The benefit of this approach is that it would allow for the collection of responses regarding a higher number of treatments and outcomes and, therefore, generate a finer-grained understanding of highly individualised perspectives on future health scenarios.[vivii]

A meaningful way to then refine an understanding of the respondent's position would be to complement such sliding scales with ranking questions where respondents prioritise general goals of care and indicate the relative importance of certain personal values. This can become a reference point to better understand the profile of decision makers and serves as another valuable input for making judgements when the elicited scenario does not directly correspond to the real-life clinical scenario that presents. It could also function as a reflection point, enriching respondents' consideration of their choices, and possibly serve to validate the consistency of their stated preferences with their stated goals and values. Finally, the use of an algorithm to predict patients' preferences may give rise to fairness concerns. Although the existence of demographic disparities is not necessarily indicative of a discrimination[32] and stated goal of care preferences can be the result of different cultural norms and belief systems,[33] future research should investigate whether indicators of harmful demographic disparities, such as financial indicators and their proxies, may impact the process of stating goal of care preferences.[viii] To do so, they may consider collecting feedback on the role of financial considerations during data collection via questionnaires and interviews, and applying fairness criteria while training the model at the life-cycle 'modelling' step.[32]

[vi]The AI would then compute the likelihoods for each treatment and outcome.

[vii]This approach holds a few additional advantages. The task of selecting min/max likelihood on a sliding scale allows the expression of clinically appropriate and outcome-specific preferences. Thus, it offers a narrowly confined target variable allowing the collection of preferences at a very granular level (as opposed to, eg, binary target variables). Moreover, it needs no human annotation (as opposed to collecting free text).

[viii]For example, financially disadvantaged participants to data collection may repeatedly state refusals of medical treatments (eg, indicating min/max likelihoods close to 100%/0%), even in presence of favourable functional outcome levels in the proposed scenarios, due to the fear of becoming a burden for their families in case of prolonged medical treatments and post-treatment interventions.

Bias in future-self forecasting

In the context of predicting patients' preferred treatments, designers must address a peculiar source of bias: the one that emerges in the very act of declaring preferences. Research shows that biases distort people's ability to accurately recall past experiences and realistically forecast future scenarios, leading to errors in decision-making.[34] Even competent patients with high levels of health literacy often have difficulty realistically anticipating how they would respond to future disability.[34] The main example is impact bias,[35] produced by focalism, that is, focusing excessively on what is lost (eg, as a consequence of a disability), and by immune neglect, that is, the underestimation of coping strategies. As a result, people are poor safeguards of their future well-being: their biases affect the formation of realistic beliefs about their future quality of life and lower their ability to make meaningful and accurate preference predictions. Moreover, preferences tend to evolve over time, affected by one's physical and psychological condition. Although studies have explored mitigating such bias in medical decision-making,[36] more research is needed to develop methods for assessing and managing bias in predicting future-self preferences. The same is true for further decisional challenges such as decision-making under conditions of uncertainty and probabilistic thinking.

To address these limitations, researchers have advocated for the use of technology to improve people's decisional capacity when completing advance directives.[23 37] Existing efforts include multimedia tools, low-health-literacy multilanguage printouts and 'e-planning' solutions, that is, web-based applications.[5 23] Some authors have recently remarked that, more broadly, digital technologies have the potential to aid informed advanced care planning by supporting the four components necessary for decisional capacity[38], that is, understanding, appreciation, reasoning and communication. Using digital technology (eg, a browser-based app), those completing advance directives can be guided towards supportive elements such as video accounts from others' lived experiences,[39] forums for peer exchange and interactive thought exercises[38]. By engaging with this interactive material, users can familiarise themselves with future scenarios, gaining insights into how they themselves might adapt and respond to changing circumstances[38]. This interactive material might improve user's ability to vividly project themselves into future healthcare scenarios, leading to a more likely hypothetical picture of the respondents future care goals and preferences[38]. These are concrete ways that digital technology can be applied to enhance future forecasting. Moreover, this support of the process of preference elicitation would both improve standard advance directives and set a foundation for gathering good quality data to train preference predicting algorithms.

Moreover, the use of digital tools would ensure the scalability of data collection and allow reaching far more participants than traditional methods.[5] Even if biases and other challenges were not completely eliminated, a significant improvement as compared with the status quo would–in the absence of other disadvantages–be enough to justify the use of AI-based advanced decision support.

Modelling

Multiplicity in modelling

An open question is the number of ML models required to implement an AI to predict patient preferences. The literature does not investigate this point, instead referring to 'the' algorithm that would predict preferences.[10] However, the specificities of data collection and ML modelling suggest a more complex picture. In fact, designers would need to train one ML model per

treatment scenario, either to classify the likelihood of agreeing with the proposed treatment in the scenario or to compute the maximum (or minimum) likelihood of outcome that would make the patient consent to the proposed treatment. This results in the maintenance of tens of ML pipelines, each one possibly characterised by different modelling procedures. To solve this challenge, designers may try reducing the number of treatment scenarios to be collected depending on the different levels of ML model performance, aggregating 'contiguous' scenarios while preserving clinical relevance,[ix] or ranking them and then repeating the modelling procedures. As a result, the AI would predict only a selection of all possible treatments yet retain its relevance for shared decision-making and system manageability across the different steps of its life-cycle.

Technical bias and transparency

The technical choices that designers make during modelling are another potential source of bias. Bias can be introduced through the choice of ML models, the selection of performance measures, and the methodologies to reduce class imbalance in data, that is, the presence of an unequal distribution of classes (eg, 'intervention: yes' and 'intervention: no') per each treatment scenario.[x] These design choices may be dictated by theoretical reasons, such as different accuracy-complexity trade-offs, or pragmatic reasons, such as regulatory requirements. As a result, it becomes necessary to align the implementation of technical choices with the epistemic and ethical expectations of AIs. To do so, designers can draw on recent accounts of AI transparency, such as 'design transparency',[40] that support the disclosure of 'the standards, norms and goal that were implemented in the system',[40] how those norms were translated into technical requirements, the choice of performance measures and the consistency of these choices throughout model retraining.

Explainability in modelling

Without appropriate 'epistemic guarantees' during ML modelling, clinicians may feel compelled to thoroughly validate each AI prediction during decision-making, undermining the benefits and feasibility of AI support. Relatedly, clinicians may be concerned that they must accept the opacity of the algorithms and accept an inability to satisfactorily trace the justification for the given suggestion. The challenge posed by explainability emerges at multiple stages: when data scientists and clinicians aim to validate the ML models through an understanding of the underlying logic of outcomes; when they assess performance; and when they identify patterns, including errors. Although explainability has been evoked as a requirement for algorithms that predict patient preference,[8 9] no explicit recommendations have been made.

Here, we propose a two-step approach to address the need for explainability during 'Modelling'. First, as the overall input needed for training a preference predictor model is relatively small compared with other deep learning use cases in healthcare,[13] designers should start by training models that are interpretable by design. Examples include generalised additive methods and decision trees.[41] Second, while testing more complex models to improve performance and better capture patterns in data, they should rely on methods such as Shapley values[42] to compute feature importance scores and provide a degree of protection against spurious correlations. Moreover, they should use post hoc interpretability methods, such as counterfactual explanations,[43] to explain any given conclusion. Counterfactuals elucidate a suggestion (ie, avoid subjecting the patient to a certain procedure) by providing an alternative scenario that describes the conditions under which the alternative prediction would have emerged. Counterfactuals are algorithmically easy to generate and similar to how clinicians often communicate. For example, in the ICU, a counterfactual may explain the suggestion against the placement of a shunt to drain cerebrospinal fluid in a 67-year-old incapacitated patient that would have a 90% likelihood of leading to moderate cognitive disability:

> If the patient had been 60 years old and the likelihood of moderate cognitive disability 75%, then he might likely have chosen shunt placement instead of palliative care, ceteris paribus

However, although explanations, such as counterfactuals, are a valuable tool in the 'Modelling' step of the AI system, they are neither necessary nor sufficient to justify the ethically sound use of AI, foster trust in it and increase its acceptance rates.[17 44] Therefore, designers should consider their use only in addition to extensive performance evaluation of the AI and its empirical testing with stakeholders.

Performance evaluation: AI versus surrogates

In the 'Modelling' step, the accuracy of the proposed AI is tested against the baseline of surrogates' performance in predicting preferences of their loved ones. The importance of this comparison is often emphasised in the literature on algorithmic prediction of patient preferences[1 10 26]; in fact, the key assumption is that a high performing AI will promote goal concordant care better than stand-in decision-makers. However, this assumption must be tested with appropriate methods.[xi] Surrogates and next-of-kin may have limited knowledge of the patient's preferences; their perceptions of the patient's values and goals may be biased; and the knowledge they do have may not directly relate to the situation at hand. The decision-making of surrogates and next-of-kin may be guided by salient memories and their own life views and additionally shaped by interactions with other loved ones and stakeholders.

In summary, surrogates' predictions are the result of the accrual of disparate information over time in processes that cannot be a priori modelled or appropriately elicited in a laboratory setting. Therefore, we argue, the epistemic value of an AI versus surrogates performance comparison is to provide a prima facie evidence of the accuracy of the algorithm and should not be overestimated. It should be complemented by other performance assessment routines and allow informing refinements of the 'Data Collection' and 'Modelling' steps based on surrogates' feedback and decision-making procedures.

System design

Embedding values in system design

The lack of a structured body of HCI research specific for AI systems, especially for clinical applications, affects the development of an AI to predict patient preferences. AI systems often fail to move to practice since clinicians are reluctant to use them despite good in vitro performance.[45] Moreover, when implemented in clinical practice, the subpar integration of these systems into the

[ix] Such as those describing the same treatment and outcome, but different (eg, 'high', 'medium' and 'low') likelihoods or durations.

[x] For example, during data collection, only 2% of respondents could agree on undergoing cardiopulmonary resuscitation procedures in the case of a high risk of cognitive deficit and low likelihood of survivability. Class imbalance seems reasonable in presence of particularly unfavourable treatment scenarios.

[xi] To compare the accuracy of the AI versus the stand-in decision maker at predicting goal of care preferences, one can adapt the empirical protocols of the studies where only the accuracy of surrogates is tested[24] by letting the AI compute predictions independently of the surrogates and comparing error rates with respect to the stated preferences of patients.

workflow impedes clinician motivation to use them.[46] Therefore, designers can promote a sociotechnical-aware design of clinical AIs supporting decision-making by implementing value-sensitive and user-centred design principles. Value-sensitive design accounts for human values throughout the design process, allowing developers to translate a set of values into system requirements.[47] Here, the values of interest are the embedded values, or 'values that have been intentionally, and successfully, embedded in an AI system by its designers'.[19] For example, value-sensitive design can be used to introduce requirements that translate the values promoted by the European Union guidelines for trustworthy AI,[48] such as transparency and explainability. In the case of AIs that predict patient preferences, relevant values would include respect for patient autonomy, non-maleficence, preservation of clinician agency, promotion of shared decision-making, timeliness and commitment to providing goal concordant care. The translation of values in a set of compatible system requirements can be facilitated by using structured questionnaires and design workshops involving clinicians, surrogates and, possibly, patients enrolled from studies on retrospective agreement (see the 'Evaluation' section).

What is a good system design?

Finally, user-centred design offers the possibility of developing systems based on the characteristics and tasks of their users.[49] Designers should leverage user-centred design of clinical AIs to ensure high degrees of usability and robustness, considering different interfaces (eg, tablets, computer monitors, interactive screens); the particularities of the clinicians' working schedule; clinicians' comfort with technology; and the characteristics of the location for the consultation with loved ones. Moreover, if the AI is directly used with loved ones, the display of patient information and treatment predictions should allow the assessment of the system trustworthiness by heterogeneous populations of stakeholders with diverse languages and belief-systems. Therefore, designers should consider multilingual interfaces, using vocal support functionalities or simplified content and interactions in the case of specific decision-maker impairments. Finally, the integration of AI into clinical workflows should support timely decision-making and the collection of patient data after the choice of treatment, such as in the case of studies on retrospective agreement (see the 'Evaluation' section).

Deployment

Integration into clinical shared decision-making

Another key consideration is how a decision-making process involving clinicians, loved ones and an AI might unfold. On the one hand, the AI may be implemented in a workflow to provide predictions only to clinicians. As a pilot study has shown, healthcare professionals may be quite open to the use of an AI to inform what treatment a patient may likely want.[9] In this scenario, loved ones would not be presented with an algorithmic prediction of preferred treatments but, rather, with a suggestion by a 'clinicians+AI' dyad.[17] Similar to the 'modelling' step, explainability methods would support clinicians in forming a mental model of the AI prediction to then take into account its suggestion as part of their preparation for consulting with loved ones.[xii] Consideration would need to be given to how the best clinicians should, if at all, present the information provided by the AI. An AI-based decision-support designed primarily for clinician use may be especially useful when no surrogates are available or surrogates are unable or unwilling to assume their role.

In another scenario, both clinicians and stand-in decision makers could consult the AI and its predictions, either individually or in a joint process. Surrogates or next-of-kin would be given the possibility to interact with the AI to improve their understanding of its prediction through explanations that might at the same time foster their trust in the system. Recent empirical evidence supports this possibility, suggesting that surrogates may respond positively to the support of an AI to predict the preferences of their loved ones.[12] Stand-in decision makers may benefit from having a guide in the deliberation process and feel less isolated in decision-making; however, AI should be limited to the role of a support tool, since surrogates rightly emphasise their unique role and authority to decide which treatments their loved ones would want accepted or refused.[12]

One concern is that surrogates or next-of-kin may disagree with the AI and its predictions, perceiving the AI as challenging their decision-making or undermining their confidence.[26] Potential conflict between surrogates and an AI tool would undermine the key hypothesis that 'if a[n] [AI] can accurately predict patient treatment preferences, it also may reduce surrogate distress'.[11] Therefore, conflicts arise when accurate predictions on historical data[xiii] are not sufficient for the surrogates to trust the AI prediction, leading, arguably, to surrogates' stress in decision-making. In summary, surrogates and next-of-kin may approve of having a guide in the form of an AI-based PPP, and this support may improve efforts to provide goal concordant goal.[12] However, outcomes perceived as surprising, puzzling or incorrect may still result in discord. Such conflict may undermine the value of using the AI in clinical practice.[26]

This said, we argue that the emergence of conflicts between those close to the patient and an AI-based PPP is to be managed, not pre-emptively avoided. Conflicts can also emerge between loved ones and clinicians when clinicians have a position on what is in the patient's best interest. In this case, disagreements are often resolved through consultation–possibly involving a clinical ethicist–and discussion of the different options at stake. Similarly, we argue, when there is disagreement with a suggestion generated by an AI, it becomes relevant to support understanding of what underlies the AI prediction, creating space to reflect on potential biases while minimising distress from lack of appropriate information, inadequate explanations, or inefficient communication.

We believe it remains an open question for now which scenarios of AI-assisted decision-making are ethically sound to implement: should such AI be reserved for aiding clinician decision-making or is there a potential application for direct use by loved ones? Although the literature agrees on the supporting role of an AI predicting patient preferences,[1 9 12] it is not yet clear who should draw on its predictions, at what point and why. We argue that, at the time of writing, more empirical research is needed to better understand the interactions of loved ones, clinicians and such an AI system. However, in general, we think access should be provided to both clinicians and loved ones with due explanations and as desired. Also, the AI-assisted decision-making process may benefit from professional moderation through a clinical ethics consultant, particularly if loved ones are uncomfortable with the AI suggestions. In

[xii] Although the provision of epistemic guarantees is considered to be necessary to hold justified beliefs on the trustworthiness of an AI,[44] the nature of these guarantees is currently under discussion. Specifically, it is debated whether explainability methods are an example of such guarantees.[17 44]

[xiii] Accuracy is a property of an AI that is measured on a given dataset during a performance evaluation of the system. It refers to a performance on historical data and cannot be computed for a given prediction. Therefore, expressions such as 'the PPP increases surrogates' predictive accuracy'[12] and cognates refer to a standard that is achieved before decision-making in clinical practice. As a result, during decision-making, accuracy may support loved ones in assessing the trustworthiness of the AI.

addition, an important ethical consideration to probe is how clinician use of such AI might augment the clinician's authority and affect the power balance in decision-making between clinicians and surrogates or next-of-kin. On the one hand, there is concern that lack of explainability and insufficiently transparent AI suggestions may erode trust and detract from the value of clinician input.[17 44] On the other hand, there is concern that the perceived credibility of the AI suggestion may discourage clinicians from applying their own judgement or may unduly weigh decision-making in the direction of the augmented AI-clinician recommendation diminishing loved one's sense of empowered involvement in joint decision-making.[50] Therefore, the way in which AI recommendations might influence the weight of various stakeholders' input in shared decision-making is something to be considered carefully and further explored.

Finally, should AI-assisted decision-making in the future prove to be vastly superior to unaided decision-making and become the gold standard, the voluntariness of its use by healthcare professional might need to be revisited. In that case, the AI might hold an assumed role as quasi-surrogate, and patients would need to explicitly state in an advance directive their preference to decline its use.

Evaluation

Performance evaluation: goal concordant care

Retrospective agreement studies offer the possibility of testing the performance of the proposed AI in supporting goal concordant care. This performance validation process makes use of qualitative and quantitative metrics[41] and is grounded by studies set in realistic settings, as opposed to the validation in the 'Modelling' step.

Considering the ICU use case again, two distinct types of studies on retrospective agreement can be used to evaluate performance of the proposed AI. In the first study, the predictions computed at relevant ICU decision-making moments may be compared with the decisions made by clinicians for a sample of patients following discharge. This would minimise risk during the training phase by presenting clinicians with the AI predictions only after patients are discharged. Clinicians could then review all cases by means of the patient charts, describing their decision-making processes (with or without surrogate decision-makers, available ADs, etc), and comparing their outcomes with the preferences predicted by the clinical AI at each selected decision-making moment. This type of study evaluates the rationale behind the clinical AI predictions as it compares to clinicians' judgement. It could be run when the AI is newly deployed in the clinical workflow, periodically to monitor the accuracy of the system, or to train clinicians in AI-assisted decision-making on selected ethically sensitive patient cases.

A second type of study on retrospective agreement may involve patients and their loved ones in a period 6–12 months following discharge from the ICU. Questionnaires and structured phone interviews could explore and quantify the extent of retrospective agreement to intensive care, which is the stated approval of the applied treatments, depending on functional outcome as measured by, for example, the modified Rankin Scale (mRS) and patient satisfaction.[51] Then, to assess AI performance, one could analyse how functional outcome and satisfaction influence retrospective agreement to intensive care qualitatively and identify predisposing factors (eg, demographic and clinical data, such as preadmission mRS) for retrospective agreement to intensive care.[51] This type of study may allow comparing the decisions that the AI suggested with those that have been made for all patients. As a result, one could analyse for which cases, following the AI predictions, it would have resulted in higher rates of retrospective agreement to intensive care, as well as satisfaction.

These evaluations of the AI would contribute to a better understanding of the potential of these systems to foster goal concordant care, providing a degree of protection against violations of patient autonomy, avoiding preference misdiagnoses and prointervention treatment biases that may not correspond with patients' interests.

Performative prediction

Data generated in the studies on retrospective agreement can be collected and used for the modelling of the (updated) ML models predicting patients' preferred treatments. To do so, the personal-level information, the preferred treatments and the feedback from the interviews of patients following an ICU stay can be added to the existing database (or overwrite existing data points). Therefore, the deployment of an AI predicting patient preferences allows new data to be generated via studies on retrospective agreement and these data, in turn, affect the update procedures of the AI, namely the retraining of the ML models that predict the preferred treatments. This phenomenon is an example of 'performative prediction', that is, the occurrence of a shift in the distribution of patients' data that results from the deployment of an ML model.[52] Performative prediction breaks the common assumption in ML modelling that data distributions are somehow static or that their change over time depends on exogenous causes–such as a pandemic–providing an endogenous cause for the shift of the patient data distribution: the deployment and use of ML models to assist decision-making.[xiv]

As a result, if an AI to predict patient preferences is incorporated into use, designers have to acknowledge that the choice of which patients to interview in studies on retrospective agreement, how to proceed with data collection, and the methods of analysis of their data all potentially introduce bias[xv] that affects the 'Data Collection' and 'Modelling' steps due to performative prediction. In particular, this process may contribute to reinforce class imbalance. Management of this occurrence suggests that the sampling strategies and experimental protocols of studies on retrospective agreement become part of the documentation supporting the transparency of the AI system and its maintenance. Finally, model retraining routines have to be scheduled to cope with the patient data distribution shift over time.[xvi]

CONCLUSIONS

The integration into clinical practice of algorithms that predict care preferences, particularly when involving advanced AI, is a complex and interdisciplinary exercise that draws on ethicists, computer scientists, designers, clinicians, loved ones and patients. In this work, we have proposed a new approach to organise this endeavour, suggesting actionable recommendations to hopefully fill a gap that has persisted since the early days of the proposed PPP. At the basis of our approach lies the conviction that the interdisciplinary nature of all stages of the design of AIs to predict patient preferences should be harnessed yet organised appropriately, moving beyond taking either theoretical or pragmatic questions in isolation. To do so, we

[xiv] An exogenous source of shift of the patient data distribution could be the COVID-19 pandemic. In fact, during a pandemic, an increasing number of people face the possibility of hospitalisation—including incapacitation—and become more knowledgeable on the risks and consequences of medical treatments through media. This may affect the process of elicitation of goal of care preferences, as noted by Auriemma *et al*.[53]

[xv] This bias may manifest, for example, in collecting data mostly from patients who survived at after discharge or from those with low levels of cognitive deficit, so they can actively participate in the study.

[xvi] Model retraining can be performed periodically, that is, once per year, or ad hoc depending on cues on the shift of the data distribution (eg, analyses of feature and target variable distributions over time) and other constraints (eg, the resources needed for studies on retrospective agreement, fairness or explainability-driven requirements).

have proposed to consider AIs as sociotechnical systems, considering their life-cycle as a guide for our discussions. This theory-driven yet application-oriented approach has allowed us to discuss existing and new challenges for implementation of AI preference predictors into clinical workflows. Our attempt aims to ignite a critical discussion on how to best shape the use, evaluation and continuous improvement of algorithms assisting ethically challenging decisions in clinical practice. Hopefully, this discussion will lead to the generation of sought-after evidence regarding the performance of AI and its potential supporting role of human decision-making around most likely preferred treatments.

Correction notice This article has been corrected since it was first published. The open access licence has been updated to CC BY. 17th May 2023.

Contributors NB-A originally proposed the line of research. AF drafted the manuscript and is the guarantor of this work. SG and NB-A provided important intellectual inputs to finalise the manuscript. All authors contributed to and approved the final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Andrea Ferrario http://orcid.org/0000-0001-9968-9474
Sophie Gloeckler http://orcid.org/0000-0002-7658-823X
Nikola Biller-Andorno http://orcid.org/0000-0001-7661-1324

REFERENCES

1 Rid A, Wendler D. Can we improve treatment decision-making for incapacitated patients? *Hastings Cent Rep* 2010;40(5):36–45.
2 Kiker WA, Rutz Voumard R, Andrews LIB, *et al*. Assessment of discordance between physicians and family members regarding prognosis in patients with severe acute brain injury. *JAMA Netw Open* 2021;4(10):e2128991.
3 Rutz Voumard R, Kiker WA, Dugger KM, *et al*. Adapting to a new normal after severe acute brain injury: an observational cohort using a sequential explanatory design. *Crit Care Med* 2021;49(8):1322–32.
4 Yadav KN, Gabler NB, Cooney E, *et al*. Approximately one in three us adults completes any type of advance directive for end-of-life care. *Health Aff* 2017;36(7):1244–51.
5 Sudore RL, Schillinger D, Katen MT, *et al*. Engaging diverse English- and Spanish-speaking older adults in advance care planning: the prepare randomized clinical trial. *JAMA Intern Med* 2018;178(12):1616–25.
6 Spalding R. Accuracy in surrogate end-of-life medical decision-making: a critical review. *Appl Psychol Health Well Being* 2021;13(1):3–33.
7 Buchanan AE, BA E, Brock DW. *Deciding for others: the ethics of surrogate decision making*. Cambridge University Press, 1989.
8 Biller-Andorno N, Biller A. Algorithm-Aided Prediction of Patient Preferences - An Ethics Sneak Peek. *N Engl J Med* 2019;381(15):1480–5.
9 Biller-Andorno N, Ferrario A, Joebges S, *et al*. Ai support for ethical decision-making around resuscitation: proceed with care. *J Med Ethics* 2022;48(3):175-183.
10 Rid A, Wendler D. Use of a patient preference predictor to help make medical decisions for incapacitated patients. *J Med Philos* 2014;39(2):104–29.
11 Wendler D, Wesley B, Pavlick M, *et al*. A new method for making treatment decisions for incapacitated patients: what do patients think about the use of a patient preference predictor? *J Med Ethics* 2016;42(4):235–41.
12 Howard D, Rivlin A, Candilis P. Surrogate perspectives on a patient preference predictor: good idea, but I should decide how it is used, 2021. Available: https://www.researchsquare.com/article/rs-480243/v1
13 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368.
14 Jardas EJ, Wasserman D, Wendler D. Autonomy-based criticisms of the patient preference predictor. *J Med Ethics* 2022;48:304–10.
15 Mainz JT. The patient preference predictor and the objection from higher-order preferences. *J Med Ethics* 2023;49:226–7.
16 Sharadin NP. Patient preference predictors and the problem of naked statistical evidence. *J Med Ethics* 2018;44(12):857–62.
17 Ferrario A, Loi M. How Explainability Contributes to Trust in AI. In: *2022 ACM conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery, 2022: 1457–66.
18 Petersen E, Potdevin Y, Mohammadi E. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Technical Challenges and Solutions. arXiv:210709546 [cs]. Available: http://arxiv.org/abs/2107.09546
19 van de Poel I, Poel vande I. Embedding values in artificial intelligence (AI) systems. *Minds Mach* 2020;30(3):385–409.
20 Hoffmann TC, Montori VM, Del Mar C. The connection between evidence-based medicine and shared decision making. *JAMA* 2014;312(13):1295–6.
21 Schiff R, Rajkumar C, Bulpitt C. Views of elderly people on living wills: interview study. *BMJ* 2000;320(7250):1640–1.
22 Kolarik RC, Arnold RM, Fischer GS, *et al*. Advance care planning. *J Gen Intern Med* 2002;17(8):618–24.
23 Austin CA, Mohottige D, Sudore RL, *et al*. Tools to promote shared decision making in serious illness: a systematic review. *JAMA Intern Med* 2015;175(7):1213–21.
24 Shalowitz DI, Garrett-Mayer E, Wendler D. How should treatment decisions be made for incapacitated patients, and why? *PLoS Med* 2007;4(3):e35.
25 Cunningham TV, Scheunemann LP, Arnold RM, *et al*. How do clinicians prepare family members for the role of surrogate decision-maker? *J Med Ethics* 2018;44(1):21–6.
26 Rid A, Wendler D. Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible? *J Med Philos* 2014;39(2):130–52.
27 Houts RM, Smucker WD, Jacobson JA, *et al*. Predicting elderly outpatients' life-sustaining treatment preferences over time: the majority rules. *Med Decis Making* 2002;22(1):39–52.
28 Smucker WD, Houts RM, Danks JH, *et al*. Modal preferences predict elderly patients' life-sustaining treatment choices as well as patients' chosen surrogates do. *Med Decis Making* 2000;20(3):271–80.
29 Lamanna C, Byrne L. Should artificial intelligence augment medical decision making? the case for an autonomy algorithm. *AMA J Ethics* 2018;20(9):902–10.
30 John SD. Messy autonomy: commentary on patient preference predictors and the problem of naked statistical evidence. *J Med Ethics* 2018;44(12):864.
31 Fried TR, Bradley EH, Towle VR. Assessment of patient preferences: integrating treatments and outcomes. *J Gerontol B Psychol Sci Soc Sci* 2002;57(6):S348–54.
32 Barocas S, Hardt M, Narayanan A. Fairness in machine learning. Available: https://fairmlbook.org
33 Sudore RL, Knight SJ, McMahan RD, *et al*. A novel website to prepare diverse older adults for decision making and advance care planning: a pilot study. *J Pain Symptom Manage* 2014;47(4):674–86.
34 Halpern J, Arnold RM. Affective forecasting: an unrecognized challenge in making serious health decisions. *J Gen Intern Med* 2008;23(10):1708–12.
35 Wilson TD, Gilbert DT. The impact bias is alive and well. *J Pers Soc Psychol* 2013;105(5):740–8.
36 Almashat S, Ayotte B, Edelstein B, *et al*. Framing effect debiasing in medical decision making. *Patient Educ Couns* 2008;71(1):102–7.
37 Biller-Andorno N, Biller A. The advance care compass- a new mechanics for digitally transforming advance directives. *Front Digit Health* 2021;3.
38 Gloeckler S, Ferrario A, Biller-Andorno N. An ethical framework for incorporating digital technology into advance directives: promoting informed advance decision making in healthcare. *Yale J Biol Med* 2022;95(3):349–53.
39 Ziebland S, Herxheimer A. How patients' experiences contribute to decision making: illustrations from DIPEx (personal experiences of health and illness). *J Nurs Manag* 2008;16(4):433–9.
40 Loi M, Ferrario A, Viganò E. Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics Inf Technol* 2021;23(3):253-263.
41 Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2021;113:103655.
42 Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in neural information processing systems 30*. Curran Associates, Inc, 2017: 4765–74.
43 Wachter S, Mittelstadt BDM, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv J L & Tech* 2018;31.
44 Durán JM, Jongsma KR. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021;47(5):329–35.
45 Elwyn G, Scholl I, Tietbohl C, *et al*. "Many miles to go …": a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Med Inform Decis Mak* 2013;13 Suppl 2(Suppl 2):S14.
46 Sittig DF, Wright A, Osheroff JA, *et al*. Grand challenges in clinical decision support. *J Biomed Inform* 2008;41(2):387–92.
47 Friedman B, Kahn PH, Borning A. Value sensitive design and information systems 2013.
48 HLEG A. Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment, 2020. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
49 Brunner J, Chuang E, Goldzweig C, *et al*. User-centered design to improve clinical decision support in primary care. *Int J Med Inform* 2017;104:56–64.

50 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46(3):205–11.
51 Kiphuth IC, Köhrmann M, Kuramatsu JB, *et al*. Retrospective agreement and consent to neurocritical care is influenced by functional outcome. *Crit Care* 2010;14(4).
52 Perdomo J, Zrnic T, Mendler-Dünner C. Performative Prediction. In: *Proceedings of the 37th International Conference on machine learning*. PMLR, 2020: 7599–609.
53 Auriemma CL, Halpern SD, Asch JM, *et al*. Completion of advance directives and documented care preferences during the coronavirus disease 2019 (COVID-19) pandemic. *JAMA Netw Open* 2020;3(7):e2015762.